Assignment I : Data Preprocessing, Handling Missing Values, Grafana

June 4, 2025

ID5003W: Industrial AI at Scale Laboratory

Total Marks:50

Submission Deadline : June 16, 2025

General Instructions

1. Assignment shall be submitted before the due date. Late submissions will not be entertained. If you cannot submit the assignment due to some reasons, please contact Dr. Alka Bhushan by email.
2. All the assignments must be the student's own work. The students are encouraged to discuss or consult friends or classmates. However, they have to submit their own work. Any malpractice will be reported to the authorities and action will be taken as per the IIT Madras rules.
3. If you find the solution in the book or article or on the website, please indicate the reference in the solution

Dataset

Each student is provided with a dataset containing three csv files : customers.csv, products.csv, and orders.csv. Each file includes missing values and corrupt records.
The customers.csv file includes nested JSON strings in fields preferences (array of strings) and address (a struct with street, city, zip).

Problem 1 : Data Ingestion and Cleaning (Marks: 10)

1. Read all the files into spark dataframe with schema inference
2. Print the size of each dataframe and schema
3. Identify the corrupt values in each file and report the number of corrupt rows and total corrupt values in each file.
4. For numerical columns, fill the missing values with mean value
5. For categorical or string columns, fill the missing values with most frequent value.

Problem 2: Data Transformation (Marks: 10)

1. Compute total spending by each customer
2. Identify top 3 products per customer by quantity
3. Create a summary dataframe with
   - customer_id
   - total_orders
   - total_spent
   - top_3_product_ids
4. Display top 50 customers in terms of total orders.

Problem 3: Advance Analysis (Marks: 10)

1. Identify and find how many customers have anomalous behavior using window functions and statistical thresholds. Statistical threshold is mean + 3*standard_deviation.
2. Display 50 customers who shows anomalous behavior.

Problem 4: Performance Profiling (Marks: 10)

1. Run the cleaning, transformation and advance analysis with different spark configurations:
   - Executor instances : 1 or 2
   - Executor cores : 1 or 2
   - Executor memory : 512 MB or 1 GB
2. Print the spark configuration
3. Measure the run time of each important step and (use time.time()) and log the run time in a csv file.
4. Analyze the log returned after execution of the code and
   - Report the number of jobs created by the submitted code and total time taken by each job
   - Report the number of stages in each job
5. Plot the metric using matplotlib

Problem 5: Grafana (Marks: 10)

Use Grafana CSV data source to upload csv file created in Problem 4 and

- Create a pie chart to show the proportion of time spent in data ingestion, cleaning, transformation and advance analysis for each configuration.
- Creat a bar chart to show the total run time for each spark configuration.

Execution Instructions

1. Use the sample data provided in the folder to implement problem 1, problem 2 and problem 3 on google colab

2. Change the file path before submitting the job to server to run on spark cluster

3. Download the file as '.py' file

4. Change the file name in spark_job_execution.py and submit the job in spark cluster using this file.

5. Check the log file for the output

Submission Instructions

1. Submit the log file.

2. Submit a report on the output generated for each problem and plots.

3. Submit the code that you have submitted to spark cluster.